DOCUMENT RESUME

ED 365 707 TM 020 893

AUTHOR Oshima, T. C.; And Others
TITLE Evaluation of DTF and DIF in Two-Dimensional IRT.
PUB DATE Apr 93
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993). Small type throughout document.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Equations (Mathematics); *Item Bias; Item Response Theory; Mathematical Models; Reference Groups; *Research Methodology; Validity
IDENTIFIERS *Multidimensionality (Tests)

ABSTRACT

In 1992 N. S. Raju and others proposed a general procedure for assessing differential test functioning (DTF) and item bias (differential item functioning or DIF) in tests developed with unidimensional, multidimensional, or polytomous item response theory (IRT) models. The purpose of this paper is to assess the adequacy and validity of their technique in the two-dimensional IRT setup. Following Raju and others, a chi square test is described for determining whether or not an observed DTF is significantly different from zero. When an observed DTF is statistically significant, one may begin the search for items that may be causing the significant chi square. After deletion of items with a high positive compensatory DIF, the DTF index and its chi square should be recomputed. The technique is illustrated with generated 40-item two-dimensional data sets. Once the multidimensional item parameters are given and differences in item parameters between the reference and focal groups are found, the new procedure identifies DTF and DIF as expected. Five tables present analysis results. (Contains 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM020893

ED 365 707

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T.C. Oshima
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Evaluation of DTF and DIF in Two-Dimensional IRT

T. C. Oshima

Georgia State University

Nambury S. Raju

Illinois Institute of Technology

and

Claudia Flowers

Georgia State University

Paper presented at the annual meeting of

American Educational Research Association, Atlanta, April, 1993.

Running head:DTF and DIF

TM 020 893

BEST COPY A

Many researchers have suggested that educational test data, particularly achievement test data, may not always satisfy the unidimensionality assumption in unidimensional IRT (e.g., Ackerman, 1991; Traub, 1983). Nevertheless, numerous indices for deferential item functioning (DIF) currently in use are based on unidimensional models. Oshima and Miller (1992) suggested that one of the reasons for the occurrence of false positives (i.e., non-biased items identified as DIF) was due to the use of "unidimensional" DIF indices with the multidimensional data.

In recent years, the issue of bias has been explored not only at an item level, but at the test level known as differential test functioning (DTF) (Raju, van der Linden, & Fleer, 1992, Shealy & Stout, 1992). Raju et al. (1992) proposed a general procedure for assessing DTF and item bias or DIF in tests developed with unidimensional, multidimensional, or polychotomous IRT models. Although some of the existing DIF indices (e.g., simultaneous item bias test (SIBTEST) developed by Shealy & Stout, in press) are theoretically expandable to handle multidimensional traits, Raju et al.'s technique is the first multidimensional-IRT bias index which can be used with test data that are meant to be multidimensional.

Raju et al. (1992) offered an empirical demonstration of their technique only in the unidimensional case. The purpose of this research is to empirically assess the adequacy and validity of their technique in the two dimensional IRT setup. A brief description of Raju et al.'s technique is given below.

VAILABLE

DTF and DIF in Two-Dimensional IRT

In Reckase's two-dimensional, two-parameter logistic model (M2PL) (Reckase, 1986), the probability of success on item i for subject s can be written as

$$P_i(\theta_{1S}, \theta_{2S}) = \frac{e^{(a_{i1}\theta_{1S} + a_{i2}\theta_{2S} + d_i)}}{1 + e^{(a_{i1}\theta_{1S} + a_{i2}\theta_{2S} + d_i)}} \qquad (1)$$

where a_{i1} and a_{i2} are the two item discrimination parameters associated with the two underlying dimensions (θ_1 and θ_2) and d_i is the difficulty parameter. θ_{js} is the theta for subject s on dimension j. Let the test consist of k items and have one set of item parameters for each of two groups (Reference Group and Focal Group). Let us also assume that the two sets of item parameters are on a common scale. Now, let $P_{iR}(\theta_{1R}, \theta_{2R})$ represent the probability of success on item i for examinee s as if he/she were a member of the Reference Group; similarly, let $P_{iF}(\theta_{1F}, \theta_{2F})$ represent the probability of success for the same examinee on the same item as if he/she were a member of the Focal Group. If an item is functioning differently in the two groups, P_{iR} and P_{iF} would be different for a given examinee.

Differential Test Functioning

Within IRT, an examinee's true score can be expressed as

$$T_S = \sum_{i=1}^{k} P_i(\theta_{1S}, \theta_{2S}) \qquad (2)$$

In the present setup, each examinee will have two true scores, one for being a member of the Focal Group (T_{sF}) and the other for being

a member of the Reference Group (T_{*R}). If T_{*F} and T_{*R} are equal for an examinee, then the examinee's true score is independent of group membership. The greater the difference between T_{*F} and T_{*R}, the greater the DTF. Therefore, according to Raju et al., an overall measure of unfairness or DTF across examinees may be defined as

$$DTF = \mathcal{E}_F (T_{*F} - T_{*R})^2 = \mathcal{E}_F D_S^2 = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 \qquad (3)$$

where the expectation ($\mathcal{E}$) can be taken over the Focal Group, $D_* = T_{*F} - T_{*R}$, and μ_{TF} and μ_{TR} represent the mean true score of examinees in the Focal Group and the mean true score for the same examinees considered as members of the Reference Group, respectively.

Differential Item Functioning

In view of Equation 1 and 2, Equation 3 can be rewritten as

$$DTF = \mathcal{E}[\ (\sum_{i=1}^{k} d_{i*})^2\ ] \qquad (4)$$

where $d_{i*} = P_{iF}(\theta_*) - P_{iR}(\theta_*)$. Equation 4 can also be written as

$$DTF = \sum_{i=1}^{k} [Cov(d_i, D) + \mu_{di}\mu_D] \qquad (5)$$

where $Cov(d_i, D)$ is the covariance between the difference in item probabilities for item i (d_i) and the difference between the two true scores (D), and μ_{di} and μ_D are the mean of d_{i*} and D_*, respectively. Differential functioning at the item level may now be defined as

$$DIF_i = \sigma(d_i, D) = Cov(d_i, D) + \mu_{d_i}\mu_D \qquad (6)$$

This definition of DIF will hereafter be referred to as the compensatory DIF (C-DIF) to distinguish it from the non-compensatory DIF (NC-DIF) to be defined later. Combining Equations 5 and 6, we obtain

$$DTF = \sum_{i=1}^{n} C\text{-}DIF_i \qquad (7)$$

This equation shows that the definition of DIF_i given in Equation 6 is additive for a given set of items in the sense that differential functioning at the test level is simply the sum of differential functioning at the item level, and indicates how much each item DIF contributes to the total test DTF.

If one assumes that all items in the test, other than item i, are completely unbiased, then it must be true that $d_j = 0$ for all $j \neq i$. Then Equation 6 can be rewritten as

$$DIF_i = \sigma^2_{d_i} + \mu^2_{d_i} \qquad (8)$$

which does not include information about bias from other items. This DIF index will hereafter be referred to as the non-compensatory DIF (N) index. Raju et al. (1992, 1993) show how this NC-DIF index relates to some of the currently popular, IRT-based DIF indices in the unidimensional case. The assumption underlying the NC-DIF index is not likely to be satisfied in most test development situations. Since the new definition of DIF (C-

DIF) as given in Equation 6 does not require this assumption, Raju et al. feel that it offers a more realistic measure of differential item functioning. In addition, this definition of DIF relates in a simple, additive way to a definition of total test bias (DTF) for a fixed set of items, which is not available for the current IRT-based definitions of DIF. Finally, it should be noted that an item with significant NC-DIF index may not necessarily have a significant C-DIF index. This is likely to happen whenever there are two items in a test such that one item favors the Focal group and the other item favors the Reference group. In this case, the two items may have significant NC-DIF indices and non-significant C-DIF indices. This is due to the canceling effect that is taken into account in the definitions of DTF and C-DIF indices. In general, the number of significant NC-DIF items in a test may be greater than the number of significant C-DIF items, at times substantially greater.

Significance Test for DTF

Following Raju et al. (1992, 1993), we describe a chi-square test for determining whether or not an observed DTF is significantly different from zero. Let D be distributed normally with a mean of μ_D and a finite standard deviation of σ_D. Then, under the null hypothesis that the mean of D is zero, it can be shown that $N(DTF)/\sigma^2_D$ has a chi-square distribution with N degrees of freedom. That is,

$$\chi^2_N = \frac{N(DTF)}{\sigma^2_D} \qquad (9)$$

Substituting an unbiased estimate for the variance of D, the above equation can be rewritten as

$$\chi^2_N = \frac{N(DTF)}{\hat{\sigma}^2_D} \qquad (10)$$

This chi-square test may prove useful in practice for determining whether an observed DTF index is significantly different from zero. Another statistical test which may prove useful in the present context is $\sqrt{N}(\hat{D} - \mu_D)/\hat{\sigma}_D$, which, according to the previously stated assumptions for the chi-square test, has a t distribution with N-1 degrees of freedom. Since the t and chi-square tests are likely to lead to similar conclusions when N is large, we recommend the chi-square test because of its explicit relationship to DTF as shown in Equations 9 and 10.

When an observed DTF index is statistically significant, one may begin the search for items that may be causing the significant chi-square. Since the sum of item C-DIF indices is equal to the DTF index for a given set of items, the deletion of items with high, positive C-DIF indices from the test will in general reduce the DTF index for the remaining items and may also lead to a reduced chi-square for the DTF index. Raju et al. (1993) described two procedures (Procedure A and Procedure B) for identifying biased

items using the C-DIF index and for determining the effect of their deletion on the DTF index. With both procedures, after identifying and removing items from the test, the DTF index and its chi-square should be recomputed with the remaining items. Since the value for $Cov(d_i,D)$ depends on, among other things, the number of items that are still in the test, it is recommended that a single item be identified for removal at a time and that the process be continued until the chi-square associated with the revised DTF index becomes non-significant. All deleted items may be labeled "biased" or characterized as having significant C-DIF indices. The reader is referred to Raju et al. (1993) for additional details about these two procedures.

Even though a statistical test (t or chi-square) to assess the significance of an NC-DIF index can be developed, Raju et al. (1993) did not propose one because of the current emphasis on the DTF and C-DIF measures. According to Raju et al., however, an item with an NC-DIF index greater than .003 may be regarded as having significant DIF or bias. Their rationale for the proposed decision rule is that the NC-DIF index is the mean squared error at the item level and its square-root (referred to as the root mean squared error (RMSE)) is commonly used as a measure of goodness-of-fit in other contexts (Joreskog & Sorbom, 1989). A value of .05 or greater for RMSE is typically viewed as indicating lack of fit. Since the square-root of .003 is .055, Raju et al. noted that an NC-DIF index greater than .003 may be indicative of significant DIF. This criterion may be used in

practice to identify items with significant NC-DIF.

Transformation of Item Parameters to a Common Metric

In the unidimensional case, prior to a DIF analysis, the estimated item parameters for the Focal Group (for example) are transformed to a scale underlying the estimated item parameters for the Reference Group because the item parameters from two subpopulations are only invariant up to a linear transformation (Lord, 1980). The purpose of this section is to describe a transformation procedure for the multidimensional case. Ignoring the subscript for the item, Equation 1 for a Focal Group examinee can be rewritten as

$$P_F = \frac{e^{X_F}}{1 + e^{X_F}} \tag{11}$$

where

$$X_F = a_{1F}\theta_1 + a_{2F}\theta_2 + d_F \tag{12}$$

Similarly, for a Reference Group examinee,

$$P_R = \frac{e^{X_R}}{1 + e^{X_R}} \tag{13}$$

where

$$X_R = a_{1R}\theta_1 + a_{2R}\theta_2 + d_R \tag{14}$$

It should be noted that θ_1 and θ_2 in Equations 12 and 14 are assumed to be on the same scale for the Focal and Reference Groups. Also, in Equations 12 and 14, X_F and X_R are linear

combinations of θ_1 and θ_2. It should be noted further that the item parameters for the Focal Group need not be identical to the item parameters for the Reference Group in order for the item probabilities given in Equations 11 and 13 to be equal. The item probabilities for the Focal Group and Reference Group examinees can remain the same if X_F and X_R differ by a linear transformation for a fixed set of thetas; that is, if

$$X_F = AX_R + B \tag{15}$$

In view of Equations 12 and 14, Equation 15 can be rewritten as

$$a_{1F}\theta_1 + a_{2F}\theta_2 + d_F = Aa_{1R}\theta_1 + Aa_{2R}\theta_2 + Ad_R + B \tag{16}$$

This, in turn, implies that

$$a_{1F} = Aa_{1R} \tag{17}$$

$$a_{2F} = Aa_{2R} \tag{18}$$

$$d_F = Ad_R + B \tag{19}$$

The last equation (Equation 19) is especially important because it offers a practical procedure for estimating A and B. Since d_F and d_R differ by a linear transformation, one can estimate A and B as follows.

$$A = \frac{\hat{\sigma}_{d_F}}{\hat{\sigma}_{d_R}} \tag{20}$$

$$B = \hat{\mu}_{d_F} - A\hat{\mu}_{d_R} \tag{21}$$

where σ and μ refer to the standard deviation and mean,

respectively. In practice, the multiplicative (A) and additive (B) constants may be obtained with the help of estimated d_1 and d_2. This estimation procedure was used in the current investigation. Since this transformation procedure is new, we recommend that its appropriateness and viability be evaluated further in future investigations.

Method

Design

Using a compensatory multidimensional two-parameter logistic (M2PL) model (Equation 1), 40-item two-dimensional data sets were generated. Factors of interest in this study are: (a) compensatory vs. non-compensatory DIF, (b) Uniform vs. non-uniform DIF, and (c) correlation between the two θ's. Other factors such as the number of DIF items and the magnitude of DIF were held constant.

In the two-dimensional structure, items measured both θ_1 and θ_2 throughout the test with various degrees. An example of this type of structure in practice would be a math test in which mathematical concepts are presented with various degrees of verbal expression or a test with various degrees of test-wise sensitive items or instructionally sensitive items. This structure is considered to be "essentially two-dimensional" (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991). The first factor had two levels, the non-compensatory condition and the compensatory condition. In the non-compensatory condition,

BEST COPY

all the DIF items were biased against the focal group. On the other hand, in the compensatory condition, one half of the DIF items were biased against the focal group and the other half of the DIF items were biased against the reference group. The second factor of interest also had two levels, the uniform and non-uniform DIF. The uniform DIF was defined as differences in the d parameter between the reference and the focal group. The non-uniform DIF was defined as differences between the two groups involving a_1 and/or a_2 parameter(s). Notice there are various combinations to create these non-uniform situations. The last factor of interest had two levels, r_{θ_1,θ_2} = .0, and .5.

The number of DIF items was set at four to insure "DIF" on items, because previous studies (e.g., Oshima & Miller, 1992) have shown that as the number of biased items increases, the detection rate of item bias decreases due to the violation of the dimensionality assumption. The magnitude of DIF was set at .3 on both the a parameters and the d parameter. The magnitude of .3 for the d parameter was selected considering the magnitude of DIF resulting from .5 standard deviation difference on the nuisance trait when bias was defined in the multidimensional perspective. See Ackerman (1992) for bias in the multidimensional perspective. Preceding this present study, DIF was created through additional trait(s) (θ_1 and θ_2) with the $\theta_{1,2}$ difference of .5 between the reference and the focal group. Then, resulting estimated d differences ranged around .3. The .3 difference on the a parameters was selected to coincide with other DIF studies. For

AVAILABLE

example, Kim and Cohen (1992) had the a difference (in a unidimensional IRT model) of .16 or .32.

Data Generation

Multidimensional item discrimination (MDISC) parameters were randomly chosen from a lognormal distribution with a mean 1.13 and a standard deviation .60 and multidimensional item difficulty (MID) parameters were randomly chosen from a normal distribution with a mean of 0 and a standard deviation of 1. Item directions were 0°, 30°, 45°, 60°, and 90° and they were embedded systematically throughout the test. For further details on the multidimensional item parameters and their simulation, see Oshima and Miller (1992). Ability parameters (θ_1 and θ_2) were simulated from random normal distribution with a mean of 0 and a standard deviation of 1. Then, a set of correlated θ's were generated for some conditions. The procedure to simulate correlated θ's is described elsewhere (See Oshima & Miller, 1990). The sample size of each group was 1000.

For clarity of organization, DIF items were shifted to the last items (Items 37-40). The item directions for the four DIF items were 0°, 30°, 60°, and 90°. MDISC and MID were held constant for all the DIF items (MDISC = 1.13 and MID = 0, i.e., the mean of MDISC and MID, respectively). For a condition in the non-uniform DIF, furthermore, the item direction was held constant to examine various patterns of non-uniform DIF. The selected constant for the item direction was 45°. The four of the patterns of the non-uniform DIF were selected for four DIF items.

Table 1 presents item parameters for DIF items.

Table 1 about here

Analysis

Item parameters for the generated data were calibrated using NOHARM (Fraser, 1988). For the conditions in which θ_1 and θ_2 had a 0 correlation, the correlation matrix for θs (P) was fixed as an identity matrix for both reference and focal groups. When the correlation was .5, the P matrix was fixed with .5 for off-diagonal elements and 1.0 for the diagonal elements for both reference and focal groups. Estimated item parameters were put on a common metric using a procedure described in the previous section. The multiplicative and additive constants were calculated after eliminating outliers with respect to the difference between d_1 and d_2. The DTF and DIF indices were computed using Raju et al.'s computer program.

Results and Discussion

As was presented earlier, the purpose of this paper was to evaluate Raju et al's DTF and DIF indices. It is important to keep in mind that a recovery analysis such as this study involves the evaluation of not only their DTF/DIF technique but also the performance of the calibration program. To separate the issue of the performance of the DTF/DIF technique and the performance of NOHARM in its ability to recover the item parameters, the DTF/DIF analysis was first conducted using the true parameters. This is

considered to be an optimal condition with an assumption that NOHARM recovers item parameters with no error and two sets of item parameters (one from the reference group, and another from the focal group) are put on a common scale with no error.

DIF/DTF Analysis with True Item Parameters

Results using true item parameters are shown in Table 2 for the correlation of 0. The C-DIF and NC-DIF for the non-biased items (Items 1-36) are all 0. For the uniform bias, C-DIF was .015 and NC-DIF was .004 for the biased items (Items 37-40) for the non-compensatory condition. As noted earlier the significance test for NC-DIF has not been proposed. Raju et al. (1992, 1993) suggest the cutoff value of .003 to .005. The degree of bias embedded in this study (i.e, the d difference of .3) was probably close to the lowest bound that the index can identify DIF. It is likely to have the d difference higher than .3 in practice (see Oshima & Miller, 1992). The chi-square procedure successfully eliminated the four biased items. On the other hand, for the compensatory condition, C-DIF was 0 for all the items, and the chi-square was non-significant indicating that there is no DTF, as expected. NC-DIF was .004 indicating DIF at the item level.

Table 2 about here

For the non-uniform bias, a similar trend was observed. When there was a difference on both a's and d parameters, the

BEST COPY AV

degree of bias was larger than when there was a difference only on d, indicating that Raju et al.'s indices identify not only the uniform but also non-uniform bias. The second (#2) non-compensatory condition further illustrates this point. The difference in the a parameter can be only on one of the a's or both with various directions. The NC-DIF index identified bias in all four patterns with various degrees of DIF. The chi-square procedure identified Item 40 to be deleted to reach non-significant DTF. Notice that Item 40 was the only item with the d difference. This result illustrates that only uniform bias contributed to DTF. For the non-uniform bias, although there was DIF at the item level as indicated by NC-DIF, due to the canceling effect within an item, the item did not contribute to DTF. However, this result may not be generalizable to all the non-uniform bias in general. Depending on where the item response surfaces cross in the three-dimensional space, results are likely to change. Table 3 presents the results from the conditions in which the correlation between θ_1 and θ_2 was .5 using true item parameters. The results are similar to those from Table 2.

Table 3 about here

DIF/DTF Analysis with Estimated Item Parameters

Results from estimated item parameters are shown in Tables 4 and 5 for the correlation of 0 and .5, respectively. The closer

AILABLE

these results are to those from Tables 2 and 3, the better the recovery of item parameters by NOHARM. Results from Table 4 are fairly similar to those from Table 2, indicating that the recovery of item parameters was fairly accurate and that the linkage procedure had worked well when the correlation between θ_1 and θ_2 was 0. There are, however, some minor problems. For example, in the case of uniform bias with non-compensatory items, 8 items were deleted to achieve non-significant chi-square. Of these 8 items, only four items were truly biased, thus resulting in 4 false positives. Also, in the case of non-uniform bias with non-compensatory items (#2), Item 40 did not show significant C-DIF, whereas the same item was identified as having a significant C-DIF index when true item parameters were used (Table 2).

On the other hand, results from Table 5 are not as similar to those from Table 3 as expected. For example, only one item (Item 37) was identified as having significant C-DIF in the uniform, non-compensatory condition. According to Table 3, all four truly biased items were identified as having significant C-DIF when the DTF/DIF procedure was used with true parameters. In the case of non-uniform bias with non-compensatory items, the new procedure also did poorly in the sense that it required the deletion of 14 items to achieve non-significant chi-square, of which only one item (Item 40) was truly biased. One possible explanation for the observed poor performance is the degree to which the NOHARM computer program can recover item parameters when the thetas are correlated. Another explanation is that the

new DTF/DIF is not sensitive to true DIF when the thetas are correlated somewhat. A third possible explanation is the linking procedure used in this study. All these factors should be investigated further to more accurately assess the viability of the new DTF/DIF procedure.

Tables 4 and 5 about here

In all the conditions in Tables 4 and 5, the false positive rate for NC-DIF (i.e., DIF for non-biased items 1 through 36) was within a reasonable range (not shown in Tables 4 and 5). The number of items with significant DIF using the .003 criterion ranged from 0 to 4 for the conditions in Table 4 and 1 to 4 for the conditions in Table 5. Using a more stringent criterion of .005, the number of DIF items for the non-biased items ranged from 0 to 1 for conditions in Table 4 and 0 to 2 for conditions in Table 5.

As noted earlier, the purpose of this study was to evaluate Raju et al.'s DTF/DIF indices with the two-dimensional data sets. It was demonstrated that once the multidimensional item parameters were given and there existed differences in item parameters between the reference and the focal group, the new procedure identified DTF and DIF as expected. The indices performed as well as they did with unidimensional data sets.

Now that a tool is available to detect DTF/DIF with the multidimensional test data and other tools are being developed

(W. F. Stout, personal communication, March 31, 1993), there are other problems to be solved before the indices can be put to use in practice. First, the recovery of item parameters by NOHARM or other multidimensional calibration programs needs to be investigated. With our limited data sets, it appeared that the recovery was less accurate as the correlation between the traits was introduced and the stability of recovery from sample to sample was questionable. We used the sample size of 1000 and the number of items of 40. Further studies can elaborate on the issue of stability of NOHARM as a function of the sample size and the number of items.

Second, there is a dire need for developing appropriate linkage (transformation) methods for multidimensional data sets. It appeared in this study that the chi-square technique for the DTF/DIF indices was rather sensitive to sampling fluctuations. Without an appropriate linkage method, the chi-square technique for DTF can be quite misleading. On the other hand, this study showed that the DIF index (NC-DIF) identified bias even with the presence of the problem with respect to recovery and the linkage of multiple calibrations.

Obviously, there is a great need for further research in the issue of bias with multidimensional data. The generalizability of our results is limited to the multidimensional structure and items parameters used in this study. Comparative studies involving other indices such as SIBTEST, which can be modified to multidimensional models, would be advisable.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. Applied Psychological Measurement, 15, 13-24.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Center for Behavioral Studies, The University of New England Armidale, New South Wales, Australia.

Joreskog, K. G., & Sorbom, D. (1989). LISREL 7: A guide to the program and application. Chicago, IL: SPSS Inc.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. Journal of Educational Measurement, 29, 51-66.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. Journal of Educational Measurement, 27, 273-283.

Oshima, T. C., & Miller, M. D. (1992, April). Item bias detection in the multidimensional test data. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16, 237-248.

Reckase, M. D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1992). An internal measure of test bias with applications for differential item functioning. Paper presented at the annual meeting of American Educational Research Association, San Francisco.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1993). An IRT-based internal measure of test bias with applications for differential item functioning. Manuscript submitted for publication.

Shealy, R. T., & Stout, W. F. (1992). An item response theory model for test bias. In P.W. Holland & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Erlbaum Associates.

Shealy, R. T., & Stout, W. F. (in press). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika.

Stout, W. F., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1991). DIMTEST and TESTSIM. Unpublished

manuscript.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Table 1
Item Parameters for Generating DIF Conditions

		Reference Group			Focal Groups								
					Uniform					Non-Uniform			
					Non-Compensatory			Compensatory			Non-Compensatory			Compensatory		
Item	c	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
37	0	1.13	0	0			-.3			-.3						
38	30	.98	.57	0			-.3			-.3	N/A			N/A		
39	60	.57	.98	0			-.3			.3						
40	90	0	1.13	0			-.3			.3						
or																
37	45	.80	.80	0							-.3	-.3	-.3	-.3	-.3	-.3
38	45	.80	.80	0	N/A			N/A			-.3	-.3	-.3	-.3	-.3	-.3
39	45	.80	.80	0							-.3	-.3	-.3	-.3	-.3	.3
40	45	.80	.80	0							-.3	-.3	-.3	-.3	-.3	.3

Also
-.3
-.3 -.3
-.3 .3
-.3 -.3 -.3

Note. The values under Focal Groups indicate the difference in parameters between the reference group and the focal group. Blank indicates that the parameters are the same as those in the reference group.

BEST CO

Table 2
C-DIF and NC-DIF With r = 0 Using True Item Parameters

(a) Uniform Bias

	Non-Compensatory		Compensatory	
	a_1 a_2 d		a_1 a_2 d	
		-.3		-.3
		-.3		-.3
		-.3		.3
		-.3		.3
	C-DIF	NC-DIF	C-DIF	NC-DIF
Item				
37	.015	.004	.000	.004
38	.015	.004	.000	.004
39	.015	.004	.000	.004
40	.015	.004	.000	.004
Mean*	.001		.000	
SD*	.004		.000	
DTF	.058		.001	
Chi-Square	25126.09 (p = .0000)		1000.02 (p = .4939)	
Deleted*	40-37-39-38		none	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

PY AVAILABLE

(b) Non-Uniform Bias

	Non-Compensatory		Compensatory		Non-Compensatory (β_2)	
	a_1 a_2 d		a_1 a_2 d		a_1 a_2 d	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3 -.3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 .3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 -.3 -.3	
	C-DIF	NC-DIF	C-DIF	NC-DIF	C-DIF	NC-DIF
Item						
37	.041	.010	.022	.010	.012	.003
38	.041	.010	.022	.010	.013	.005
39	.041	.010	.021	.010	.012	.007
40	.041	.010	.021	.010	.019	.010
Mean*	.004		.002		.001	
SD*	.012		.006		.004	
DTF	.164		.086		.057	
Chi-Square	1886.27 (p = .0000)		1001.27 (p = .4828)		1123.93 (p = .0037)	
Deleted*	37-38-39-40		none		40	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

Table 3
C-DIF and NC-DIF With r = .5 Using True Item Parameters

(a) Uniform Bias

	Non-Compensatory		Compensatory	
	a_1 a_2 d -.3 -.3 -.3 -.3		a_1 a_2 d -.3 -.3 .3 .3	
	C-DIF	NC-DIF	C-DIF	NC-DIF
Item				
37	.014	.004	.000	.004
38	.014	.003	.000	.003
39	.014	.003	.000	.003
40	.014	.004	.000	.004
Mean*	.001		.000	
SD*	.004		.000	
DTF	.056		.001	
Chi-Square	15438.08 (p = .0000)		1000.20 (p = .4923)	
Deleted*	37-40-38-39		none	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

(b) Non-Uniform Bias

	Non-Compensatory		Compensatory		Non-Compensatory (#2)	
	a_1 a_2 d		a_1 a_2 d		a_1 a_2 d	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3 -.3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 .3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 -.3 -.3	
	C-DIF	NC-DIF	C-DIF	NC-DIF	C-DIF	NC-DIF
Item						
37	.041	.010	.024	.010	.009	.002
38	.041	.010	.024	.010	.014	.006
39	.041	.010	.024	.010	.006	.003
40	.041	.010	.024	.010	.019	.010
Mean*	.004		.002		.001	
SD*	.012		.007		.004	
DTF	.164		.095		.048	
Chi-Square	1686.29 (p = .0000)		1000.07 (p = .4935)		1094.28 (p = .0197)	
Deleted*	37-38-39-40		none		40	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

Table 4
C-DIF and NC-DIF With r = 0 Using Estimated Item Parameters

(a) Uniform Bias

	Non-Compensatory		Compensatory	
	a_1 a_2 d		a_1 a_2 d	
	1 1 -.3		1 1 -.3	
	1 1 -.3		1 1 -.3	
	1 1 -.3		1 1 .3	
	1 1 -.3		1 1 .3	
	C-DIF	NC-DIF	C-DIF	NC-DIF
Item				
37	.076	.013	-.007	.008
38	.051	.007	.012	.006
39	.042	.005	.006	.003
40	.025	.003	.014	.007
Mean*	.011		.003	
SD*	.022		.006	
DTF	.461		.125	
Chi-Square	3550.37 (p = .0000)		1068.46 (p = .0653)	
Deleted*	40-37-38-21-31-16-39-1		none	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

(b) Non-Uniform Bias

	Non-Compensatory		Compensatory		Non-Compensatory (#2)	
	a_1 a_2 d		a_1 a_2 d		a_1 a_2 d	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3	
	-.3 -.3 -.3		-.3 -.3 -.3		-.3 -.3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 .3	
	-.3 -.3 -.3		-.3 -.3 .3		-.3 -.3 -.3	
	C-DIF	NC-DIF	C-DIF	NC-DIF	C-DIF	NC-DI
Item						
37	.061	.013	.021	.012	.000	.006
38	.052	.011	.015	.008	-.006	.005
39	.056	.013	.031	.018	.009	.009
40	.064	.011	.016	.006	-.001	.010
Mean*	.012		.003		.001	
SD*	.019		.007		.003	
DTF	.462		.104		.042	
Chi-Square	1545.78 (p = .0000)		1005.41 (p = .4460)		1024.19 (p = .2906)	
Deleted**	38-40-39-37		none		none	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
** Deleted items by Procedure B until the chi-square test reaches non-significance

BEST CO

Table 5
C-DIF and NC-DIF With r = .5 Using Estimated Item Parameters

(a) Uniform Bias

	Non-Compensatory			Compensatory		
	a_1	a_2	d	a_1	a_2	d
			-.3			-.3
			-.3			-.3
			-.3			.3
			-.3			.3
	C-DIF	NC-DIF		C-DIF	NC-DIF	
Item						
37	.006	.007		.013	.003	
38	.006	.003		-.001	.006	
39	.005	.002		.001	.001	
40	.006	.002		.007	.004	
Mean*	.001			.002		
SD*	.003			.004		
DTF	.049			.085		
Chi-Square	1118.47 (p = .0051)			1024.33 (p = .2896)		
Deleted*	37			none		

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance

PY AVAILABLE

(b) Non-Uniform Bias

	Non-Compensatory a_1 a_2 d -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3		Compensatory a_1 a_2 d -.3 -.3 -.3 -.3 -.3 -.3 -.3 -.3 .3 -.3 -.3 .3		Non-Compensatory (#2) a_1 a_2 d -.3 -.3 -.3 -.3 .3 -.3 -.3 -.3	
	C-DIF	NC-DIF	C-DIF	NC-DIF	C-DIF	NC-DIF
Item						
37	-.003	.005	.038	.011	.031	.004
38	.009	.009	.031	.011	.003	.002
39	.009	.009	.026	.013	.051	.008
40	.001	.011	.015	.009	.026	.011
Mean*	.003		.005		.010	
SD*	.005		.011		.013	
DTF	.129		.206		.369	
Chi-Square	2263.44 (p = .0000)		1023.60 (p = .2951)		1058.10 (p = .0986)	
Deleted*	40-4-36-28-5-10-7-12-8-31-6-14-19-29		none		none	

* Mean of the C-DIF for 40 items
* SD of the C-DIF for 40 items
* Deleted items by Procedure B until the chi-square test reaches non-significance